Semi Supervised Feature Learning for Tumor Growth Prediction

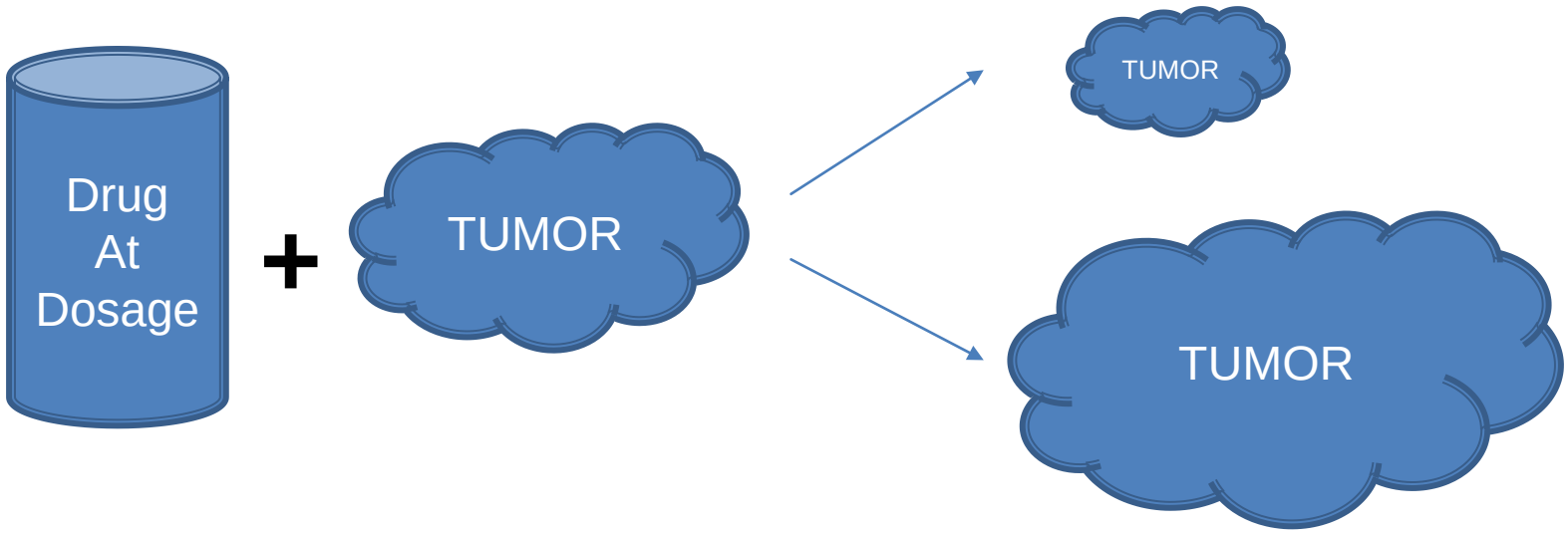
Stewart He, Jonathan Allen, Ya Ju Fan (LLNL)
Judith D Cohn (LANL)
Fangfang Xia (ANL)

LLNL-PRES-751515
This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

Prediction Problem

- Given drug features, RNASeq features, and dosage predict tumor reaction.



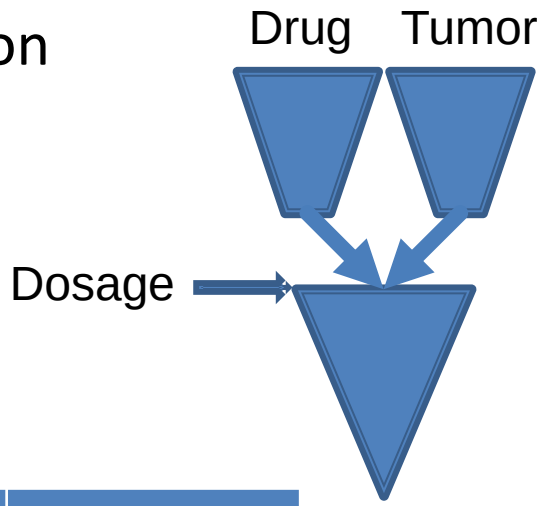
Data

- Over counts since CCLE and GDSC contain many of the same tumors under different IDs

Dataset	# drugs 3820 dims	# tumors 942 dims	Total datapoints
CCLE	24	474	84,098
CTRP	370	812	3,822,792
GDSC	247	670	1,140,574
NCI60	52641	59	18,590,413

Regression results

- Used Siamese neural network to do regression
- Perform inter-dataset test.
  - Example: Train/validate on CCLE test on CTRP
- Poor results using R^2



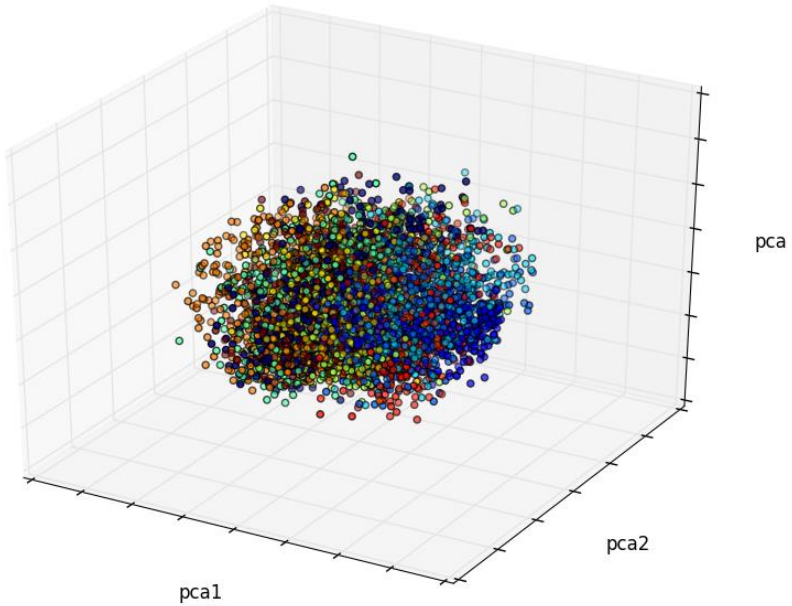
Testing-> Trianing	CCLE	CTRP	GDSC	NCI60
CCLE	.77	-.08	-.32	-.77
CTRP	.54	.81	-.12	.17
GDSC	.18	-.52	.72	-1.27
NCI60	.01	-.02	-.05	.88

Learn better features

- Very few examples of tumors
- RNASeq originally has 17k features
  - 942 landmark RNASeq genes are hand engineered
- We want better features:
  - More generalized across different types of tumors
  - Learned from unlabeled data
  - Good for regression
  - Use Autoencoders?

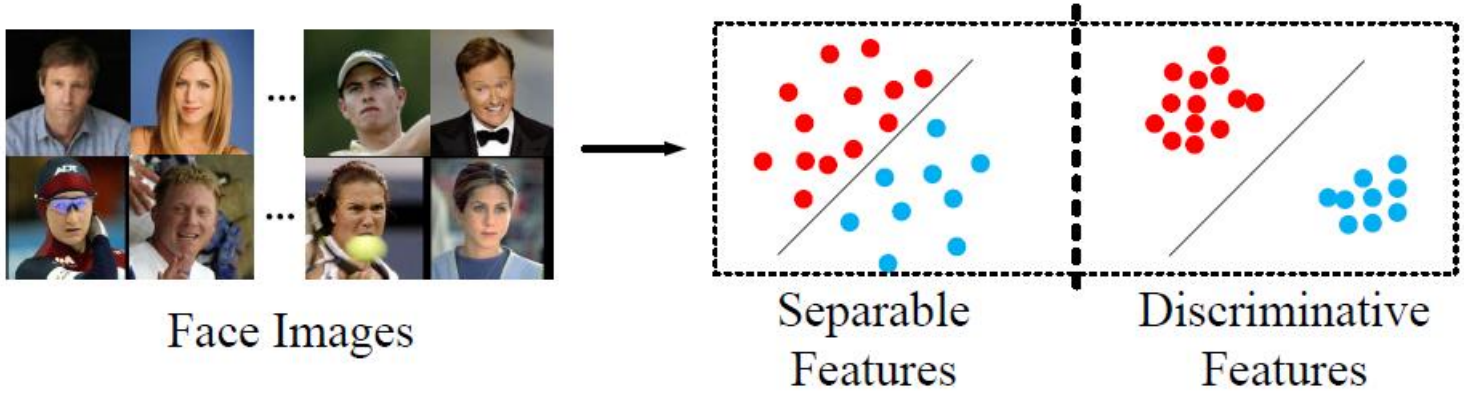
The problem with Autoencoders

- The latent space is good for reconstruction.
  - That's all the cost function cares about
  - If you're lucky they might be good for other things
- MNIST trained auto encoder
  - First 3 principal components
  - Colored by class



Modify with Center Loss

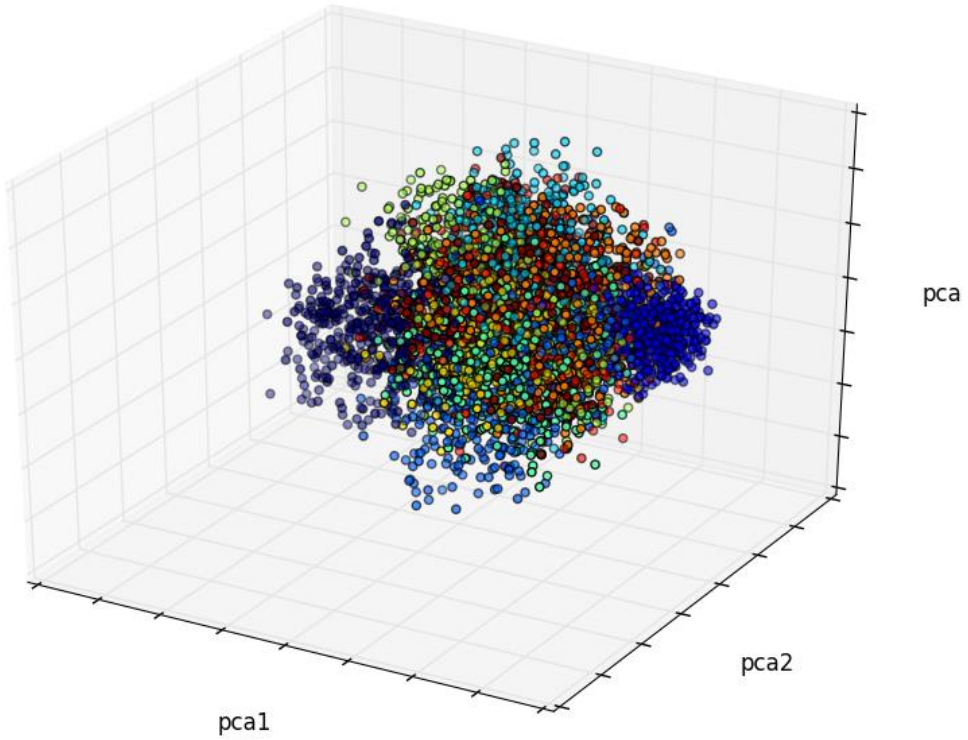
- Center Loss designed to be used for classification
  - *A Discriminative Feature Learning Approach for Deep Face Recognition* Wen et al., 2016



$$\mathcal{L}_C = \frac{1}{2}\sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{c}_{y_i}\|_2^2$$

Center Loss + Autoencoders

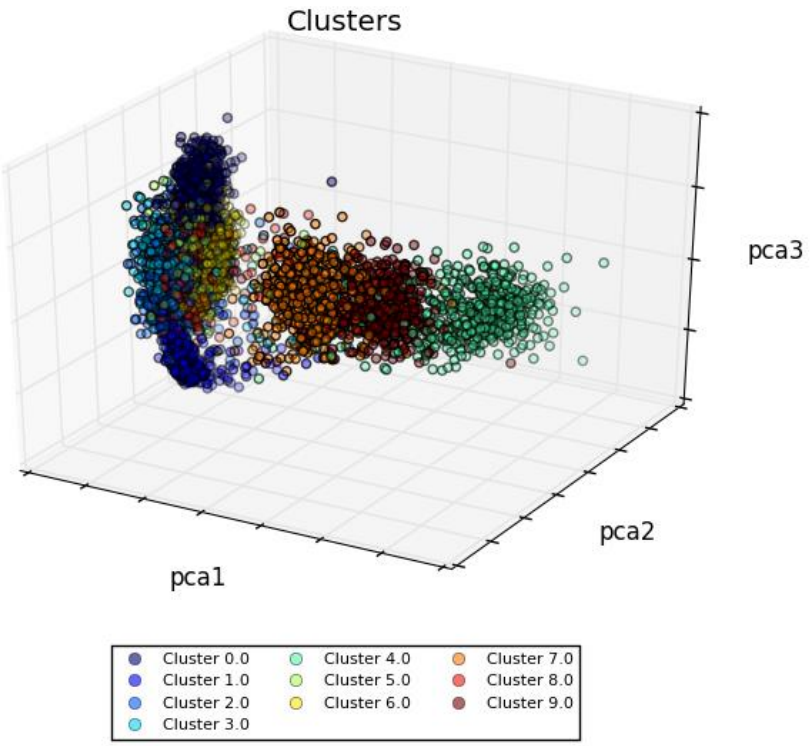
- Does not play well. MNIST example:
  - Easily falls for trivial solution



Center Loss + Center Distance

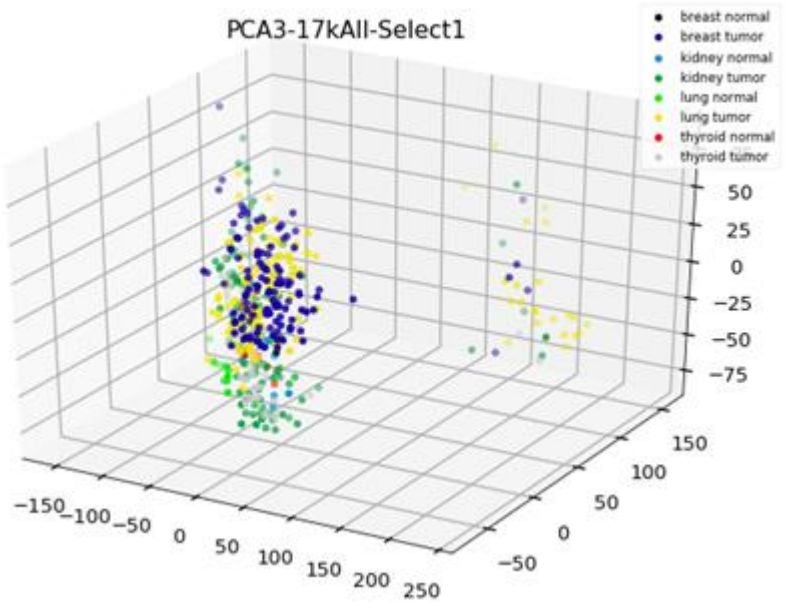
- Force the centers of classes to spread out

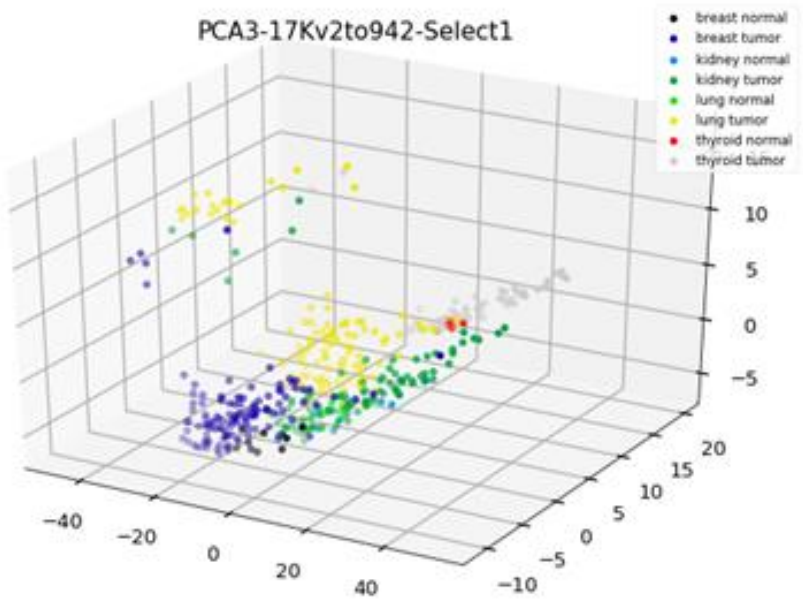
$$center\_dist = \sum_{i,j \in centers} ||c_i - c_j||^2$$



Unlabeled tumor dataset

- ~12,700 tumors 37 unique clusters (only subset shown for legibility)





Learned features applied to regression

- Modest improvement to regression results

training, testing->	ccle	ctrp	gdsc	nci60
ccle: .0001 l_r	.75	**-.325**	**-.05**	**-.3**
ctrp: .0001 l_r	**.546**	.75	**-.09**	**.24**
gdsc: .0001 l_r	**.27**	**-.41**	.76	**-.79**
nci60: .0001 l_r	**.095**	**.177**	**-.23**	.88

Acknowledgements

- Maulik Shukla - data organization/access
- Ben Mcmahon - helped Judith with clustering
- Rick Stevens - project PI


Lawrence Livermore
National Laboratory

Supplemental full image